新闻文档实体重要性排序研究*

■ 陆娜[1] 周鹏程[2] 武川[2]
[1] 海南师范大学信息科学技术学院 海口 571158 [2] 武汉大学信息管理学院 武汉 430072

摘要：[**目的/意义**]现有新闻文档实体排序研究大多以文档或实体为中心，如文本分类、实体链接等，关注实体在文本中的重要性的研究较少，本研究探讨基于重要性的新闻文档实体排序。[**方法/过程**]给定一篇文档，判断文档中实体相对文档而言的重要性，并基于此对实体进行排序。在搜狗全网新闻数据集上进行实验，并利用 NDCG 和逆序对比率两个指标对实体排序结果进行评价。[**结果/结论**]实验结果表明，基于实体频率、TF * IDF、信息熵、TextRank 等的方法以及集成方法都达到了较好的效果，基于聚集系数的方法效果一般。其中基于 TF * IDF 的方法 NDCG 值为 95.86%，是该指标下的最好结果；基于集成方法的逆序对比率值为 84.46%，是该指标下的最好结果。

关键词：新闻文档 实体重要性 实体排序
分类号：G202
DOI：10.13266/j.issn.0252-3116.2018.11.011

1 引言

实体是文档中一种特殊的语义单元，蕴含着丰富的语义信息，近年来受到越来越多的关注。目前实体相关的研究包括命名实体识别[1]、实体链接[2]、实体关系抽取[3]等，三者均属于信息抽取的范畴。命名实体识别是指识别出文档中代表命名实体的文本片段，命名实体包括7类，即人名、地名、机构名、百分比、时间、日期、货币。实体链接是指将文档中代表实体的文本片段与知识库中的特定条目相链接的过程。通常情况下命名实体识别是实体链接的第一步，即先要确定代表实体的文本片段边界，再通过实体消歧方法唯一确定实体，并链向给定的知识库，如 Wikipedia、Freebase、YAGO 等。实体关系抽取是指从非结构化的文本中抽取结构化数据，表现为主语、谓词、宾语三元组的形式，即 <Entity1，Relation，Entity2>，其中 Entity1、Entity2 为两个实体，Relation 是预定义的实体间关系。实体相关研究在信息检索、知识库构建、问答系统等领域有重要的应用价值。

实体重要性这一概念在现有研究中已受到一定关注。例如 M. Liu 等[4]在进行新闻摘要的研究中提到了关键实体的概念。然而，对实体在文档中的重要性进行专门分析的研究较少。本研究的对象是实体在文档中的重要性，即给定一篇文档，判断其中包含实体的相对重要性，并根据实体间的相对重要性对实体进行排序。传统实体排序主要分为两种，相关实体排序和面向查询的实体排序。相关实体排序是指给定一个实体和一定的限定条件，在整个实体集中寻找与给定实体之间符合限定条件的实体。面向查询的实体排序，则是面向 web 查询，返回与查询最相关的实体。此二者的搜索范围均为整个文档集中的所有实体。而本研究的任务，则是给定一篇文档，根据文档中实体对该文档的重要性进行排序，与传统的实体排序任务之间有本质区别。新闻文档是互联网中最常见的文本类型之一，相比其他类型的文档，新闻文档包含的实体数量及类型较多。基于此，本研究选择以新闻文档为例进行实体重要性排序研究。由于实体-文档间重要性关系的相关研究较少，为使本研究可行易懂，本研究将实体类型局限于人物、地点、机构3种。

本研究关注实体在文档中的重要性，通过定义实体在文档中的4个重要性等级[5]，基于实体频率、聚集

* 本文系国家自然科学基金面上项目“基于语言模型的通用实体检索建模及框架实现研究”（项目编号：71173164）和国家自然科学地区科学基金项目“基于需求社群的协商式旅游需求自动聚合方法研究”（项目编号：71762010）研究成果之一。
作者简介：陆娜（ORCID：0000-0002-5262-5286），副教授，硕士，E-mail：40303405@qq.com；周鹏程（ORCID：0000-0002-5954-6863），硕士研究生；武川（ORCID：0000-0002-8784-0808），博士研究生。
收稿日期：2018-01-02 修回日期：2018-03-02 本文起止页码：97-102 本文责任编辑：王传清

系数、TF * IDF、信息熵、TextRank 以及集成方法判别实体与实体间的相对重要性,进而对实体进行排序,通过实验比较不同的实体重要性度量方法的效果。本研究具有一定的理论意义:通过分析实体在文档中的重要性,帮助文本分析任务明确分析重点(重要实体),减小噪声(边缘实体)干扰;促进面向实体的知识组织,避免在对实体信息进行挖掘时将不相关文档纳入考虑范围。本研究可以应用于新闻分类/聚类、新闻推荐等方面。利用本研究的方法判断实体的重要性,再融合知识库中实体的有关属性,可以辅助一些文本挖掘工作。在门户网站中,如果发现用户对特定实体感兴趣,则可优先推荐包含该实体在其中的较为重要的文档,提高所推荐文档的点击率。

2 国内外研究现状

"时间""地点""人物"是新闻报道的基本要素。国内一些学者在研究新闻摘要时探讨了新闻要素重要性问题。郭艳卿等[6]提出了一种基于事件要素加权的新闻摘要提取方法,该方法将时间、地点、人物、团体机构称为新闻事件要素(本研究称其为实体),并采用中科院分词系统 ICTCLAS 识别事件要素,从多篇新闻文档中抽取事件要素组,利用事件要素出现的频率对事件要素进行加权。吴玲达等[7]进行多文档摘要时利用到了基本局部话题句群和扩展局部话题句群,其中在生成基本局部话题句群时,首先为基本新闻要素(时间、地名、人名、机构团体)赋权,其权值为新闻要素的 TF * IDF 值,然后利用聚类方法生成基本局部话题句群。这些研究在一定程度上与本研究类似,都是研究新闻文档中的实体重要性。而本研究是对给定的一篇文档,运用不同的方法判断文档中不同实体的相对重要性,并基于此对实体进行排序。

国外一些研究者提出了关键实体这一概念,并据此进行新闻摘要等研究工作。例如 K. Kiritoshi 等[8]定义了新闻文档中关键实体的概念,即新闻文档中最重要的一组实体,并用 TF * IDF 的方法识别关键实体。但是其研究问题是新闻推荐,即给定一篇新闻文档,按照与该新闻文档相关性的大小对其他新闻文档进行排序。而本研究是给定一篇文档,根据实体对该文档的重要性,对实体进行排序。M. Liu 等[9]研究了基于实体信息的新闻摘要问题,以查询词、新闻标题、句子、实体为节点,定义了标题 - 句子关系、查询词 - 句子关系、句子 - 句子关系、句子 - 实体关系等 4 种关系,建立关系图,进而利用 PageRank 算法判断句子和实体的重要性,识别重要句子和关键实体。与本研究相同,M. Liu 等[4]也关注实体在单篇文档中的重要性。其与本研究的区别在于,该研究并未对关键实体给出具体的定义,且其重点在于新闻摘要问题本身,因此运用了多种信息,包括查询词、新闻标题等。相比之下,本研究更着重研究实体重要性本身。M. Liu 等[4]提出了一种对某一时间窗口内的重要事件和关键实体进行识别的方法,从新颖性(即某实体在该时间窗口的上升趋势)和流行度(即某实体在该时间窗口内出现的频数)来判断实体的重要性程度,其更侧重于宏观分析实体在整体环境下的重要性,而本研究则着重判断实体在单篇文档中的重要性,与时间、趋势无关。

3 基于重要性的实体排序方法

新闻文档实体重要性排序是指给定一篇新闻文档,抽取其中包含的实体,并用一定的方法判断实体重要性,最后根据实体重要性大小对实体进行排序。问题的形式化定义如下:输入文档 d,且 d 包含实体 $\{e_1, e_2, \cdots, e_n\}$,其中 e_i 表示文档 d 中的第 i 个实体,输出是实体列表 $e(1) > e(2) > \cdots > e(n)$,其中 $e(i)$ 是重要性排在第 i 位的实体。假设新闻文档 $d = \{p_1, p_2, \cdots, p_n\}$,其中 p_i 是 d 的第 i 段,为了给实体重要性判断方法提供必要的输入,本研究首先对 d 进行以下预处理:对段落 p_i 进行分句处理、命名实体识别。判断实体重要性的指标有实体频率、TF * IDF、聚集系数、信息熵、TextRank 等。本研究不仅提出基于以上指标的方法,而且提出将实体频率、分布熵、实体在共现网络中的 TextRank 值等 3 个局部特征指标进行加权平均,并乘以 IDF 这一全局特征的集成方法,以解决新闻文档实体重要性排序的问题。

3.1 基于实体频率的方法

一个关于实体在文档中重要性的基本假设为:如果实体 e_i 在新闻文档 d 中出现的次数越多,那么实体 e_i 与文档 d 的相关性可能就越高,其在该文档中的重要性越高。由于要研究实体在单篇文档中的重要性,笔者认为实体的局部特征对于判断实体在单篇文档中的重要性尤为关键。统计文档中各实体出现的次数,并根据文档中实体出现的总次数进行归一化处理,见公式(1)。

$$EF_i = \frac{count_i}{\sum_{j=1}^{n} count_j} \qquad \text{公式(1)}$$

其中 $count_i$ 表示实体 e_i 在文档中出现的次数。EF(entity frequency)即实体频率,EF_i 表示实体 e_i 相对

于文档中其他实体的频率,以度量实体的重要性。

基于实体频率的方法简单直观,易于理解和解释。其局限性在于,由于某些实体的出现频数相同,在缺乏其他信息的条件下,仅利用实体频率难以对它们进行排序。这一局限性在较短的新闻文本中体现得更为明显。

3.2 基于 TF * IDF 的方法

逆文本频率(inverse document frequency, IDF)是一种衡量实体区分能力的指标。基于 TF * IDF 的方法融合实体频率和逆文本频率计算实体的重要性,见公式(2)。

$$TF*IDF_i = EF_i \cdot log\frac{N}{DF_i+1} \qquad 公式(2)$$

其中 F_i 表示实体频率,N 表示整个文档集中文档的数量,DF_i 表示实体 e_i 出现其中的文档数量。为避免出现分母为 0,这里采用了加 1 平滑方法,即对所有实体文档频数加 1。

相比 EF 这一基于单一文档的特征,IDF 是全局特征。通过综合考虑局部特征 EF 和全局特征 IDF,能够综合考虑实体在局部和全局的重要性。

3.3 基于聚集系数的方法

D. Beeferman[10] 和 T. R. Niesler[11] 在信息检索和词性识别研究中发现文档中的单词存在聚集现象,即单词之间的距离呈指数递减,且与文档相关性高的单词聚集特性更加显著。类似地,重要性程度不同的实体应具有不同的聚集特性,因此可利用反映实体聚集特性的指标判断实体的重要性。

利用实体在文档中的空间分布计算实体的聚集系数。具体地,假设实体 e_i 在文档 d 中出现的位置为 $\{p_0, p_1, \cdots, p_{n-1}, p_n, p_\{end\}\}$,这里假设实体 e_i 出现 n 次,p_i 表示 e_i 第 i 次出现的起始位置且 $p_0=0$ 表示文档的起始位置,$p_\{end\}$ 表示文档的结束位置,则实体出现的距离为 $\{p_2-p_1, p_3-p_2, \cdots, p_n-p_{n-1}\}$,实体之间的平均距离为:

$$l = \frac{1}{n}\sum_{i=1}^{n}(t_{i+1}-t_i) = \frac{p_n-p_1}{n} \qquad 公式(3)$$

距离的标准差为:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((p_{i+1}-p_i)-1)^2} \qquad 公式(4)$$

聚集系数的定义如下:

$$c = s/l \qquad 公式(5)$$

其中,l 为实体之间的平均距离,s 代表距离的标准差,c 表示聚焦系数。聚集系数越大,实体重要性越大。这是因为一般认为重要性程度高的实体在文档局部区域中频繁出现,其平均距离很小。因此本研究将计算新闻文档中所有出现的实体的聚集系数,将聚集系数高的实体视为重要实体,进行降序排列,从而实现对实体的排序。值得注意的是,文档的长度对实体的聚集系数会有显著影响。因此本研究对实体出现的位置进行归一化处理,例如实体 e_i 在文档 d 的第 i 次出现的位置为 p_i,且文档 d 的长度为 l,则实体 e_i 第 i 次出现的位置被处理为 p_i/l。这样,就可以利用归一化后的位置信息计算实体的聚集系数。

3.4 基于信息熵的方法

1949 年,香农提出了信息熵(entropy)的概念。对于同一信息源的所有可能事件,其发生的概率为 $fp_1;p_2 \cdots p_n g$,香农提出一种度量系统不确定性的指标 $s(p_1;p_2 \cdots p_n)$,公式如下:

$$s(p_1;p_2\cdots p_n) = -K\sum_{i=1}^{n} p_i\, log\, p_i \qquad 公式(6)$$

其中 K 是一个常量。信息熵具有以下性质:①$s(p_1;p_2 \cdots p_n)$是连续的;②当 $p_i=1=n$ 时,$s(p_1;p_2 \cdots p_n)$达到最大值,并且 $s(p_1;p_2 \cdots p_n)$是关于 n 的单调递增函数。

假设文档 d 由 n 部分构成,且文档的第 i 部分共包含 n_i 个实体,实体 e_i 在第 i 部分出现的频数为 $n_i(e_i)$。显然,实体在文档 d 中出现的总次数为 $\sum_{i=1}^{n} n_i$,实体 e_i 在文档中出现的总次数为 $\sum_{i=1}^{n} n_i(e_i)$,实体 e_i 在文档第 i 部分的相对频率为 $f_i(e_i)=n_i(e_i)=n_i$。

本研究定义实体 e_i 在第 i 部分分布的概率值为 $p_i(e_i) = \frac{f_i(e_i)}{\sum_{j=1}^{n} f_j(e_j)}$,则根据香农的信息熵理论,实体 e_i 分布的信息熵为:

$$S(e_i) = -\frac{1}{ln(n)}\sum_{i=1}^{n} p_i(e_i)\, ln\, p_i(e_i) \qquad 公式(7)$$

其中$\frac{1}{ln(n)}$是常量值 K,并且能保证信息熵 $S(e_i)$ 在 0－1 之间。根据段落对文档进行划分,计算文档中所有实体分布的信息熵并进行降序排列,从而对实体重要性进行排序。

3.5 基于 TextRank 的方法

对于输入文档 $d=fs_1;s_2 \cdots s_m g$,假设文档中存在 n 个实体 $fe_1;e_2 \cdots e_n g$,首先构建加权无向图 G(V;E),其中 V 表示图中节点集合,这里是指文档中的实体 e_i,E 表示图中边的集合,根据实体的共现关系确定。具体地,若实体 e_i 和 e_j 同时出现在句子 S_k 中,则二者的共

现次数加 1。统计实体之间的共现次数,用公式(8)计算 $E(e_i;e_j)$ 的权重:

$$W_{ij} = \begin{cases} \dfrac{coocur(e_i;e_j)}{\sum e_k 2Set_{e_i}\ coocur(e_i;e_k)} & coocur(e_i;e_j) > 0 \\ 0 & otherwise \end{cases}$$

公式(8)

其中 $cooccur(e_i;e_j)$ 表示实体 e_i 和 e_j 的共现次数,Set_{e_i} 表示所有与 e_i 共现过的实体集合。之所以这样计算 $E(e_i;e_j)$ 的权重,是因为经常共现的实体之间存在特定的语义关系,且共现的次数越多,实体之间存在语义相关的可能性越大。本研究利用 PageRank 算法计算节点的 PR 值,公式如下:

$$PR(e_i) = (1-d) + d \cdot \sum e_k 2Set_{e_i} W_{ik} \cdot PR(e_k)$$

公式(9)

其中 $PR(e_i)$ 表示实体 e_i 的 PageRank 值,d 表示阻尼系数,参照 PageRank 算法常用的值 d 取 0.85,W_{ik} 表示边 $E(e_i;e_k)$ 的权重。该公式符合 PageRank 的基本思想,即与实体 e_i 共现的实体越多,实体 e_i 越重要;与实体 e_i 共现的实体越重要,实体 e_i 越重要。

3.6 基于集成的方法

实体的重要性可能由多个因素决定,因此根据单一指标判断实体的重要性可能会存在偏差。例如实体的出现频率可以反映实体的重要性,但是一个实体的出现频率高并不一定意味着它对每一篇文档都很重要。此外,某些实体虽然出现的频率较高,但是这些实体可能只在文档的某一部分出现,而对文档整体的影响较小。因此实体在文档中分布也是影响实体重要性的重要因素。

本研究提出基于集成的方法,对实体的实体频率、信息熵、实体在共现网络中的 TextRank 值等 3 个局部特征指标进行加权平均,并乘以 IDF 这一全局特征。基于集成的方法计算实体重要性的公式如下:

$$c\text{-}index = (a \cdot EF + b \cdot Entropy + c \cdot TR) \cdot IDF$$

公式 (10)

其中 a,b,c 为各局部特征的权重,且有 a + b + c = 1,该权重根据启发式的方法得到。

4 新闻文档实体重要性排序实验

实验首先从搜狗实验室提供的全网新闻数据集中获取以 XML 格式存储的新闻数据,然后对其进行标记、分段落、分句以及命名识别等预处理,并将预处理结果以 XML 格式保存下来;然后通过对实体重要性等级进行定义,对随机抽取的 50 条新闻语料应用不同的排序方法进行实体重要性等级的人工标注,并且以特定的格式进行存储,见图 1。最后根据 NDGG(normalized discounted cumulative gain)以及逆序对比率两个指标对采用不同的实体重要性排序方法的结果进行评价。

4.1 新闻数据集

本研究用搜狗实验室提供的全网新闻数据进行评测[12]。该数据集收集了 2012 年 6 月 –7 月期间国内、国际、体育、社会、娱乐等 18 个频道的新闻数据,共 1 290 000多篇新闻文档。首先对新闻的正文部分进行分句、命名实体识别等预处理。文本内容是本研究的主要输入,部分不包含文本内容的新闻被过滤。

4.2 标注数据集

4.2.1 实体重要性等级定义　不同实体在文档中的重要性有所不同,包括对文档内容最为重要的实体、也有只是被简单提及 1 次的且与文档主题无关的实体。本文借鉴文献[5]中的划分,根据实体在文档中的重要性不同,把实体分为核心实体(等级为 4)、重要实体(等级为 3)、弱相关实体(等级为 2)和边缘实体(等级为 1)。重要性大小关系为:核心实体 > 重要实体 > 弱相关实体 > 边缘实体。其中核心实体是指文档围绕着这些实体展开的或者这些实体与文档的相关度明显高于其他实体;重要实体是指这些实体在新闻文档中发挥了重要作用;弱相关实体是指这些实体跟文档不是直接相关的,但是与文档的其他实体直接相关;边缘实体是指这些实体在文档中只是简单地提及,与新闻的相关程度很低。

4.2.2 标注结果　本研究从新闻语料中随机抽取了 50 篇文档,对其中的实体进行了实体重要性等级人工标注。按照特定的格式进行存储,如图 1 所示:

d4264703fd6f3339-5d2f32f06cef7000 俄罗斯:3 薛晨:4 吴鹏根:4 徐林胤:4 张希:4
波波娃:2 德国:1 科勒尔:2 澳大利亚:1 鲍登:2 帕尔默:2 巴西:1 拉里萨:2
中国队:2 波兰:1 博斯玛:2 北京:1 荷兰:1 延展:0

图 1　标注数据格式

其中 d4264703fd6f3339 – 5d2f32f06cef7000 表示文档编号,后面是文档中的实体及重要性等级。值得注意的是,本研究使用的命名实体识别工具可能会出现实体识别错误的现象,对该类实体的重要性等级判定为 0。

4.3 排序结果评价指标

本研究属于排序问题,NDCG 和逆序对比率两个指标可以对排序结果进行评价。

4.3.1 NDCG　NDCG 是一种对排序结果进行评价的

指标,该评价方法在信息检索中使用较为普遍。NDCG有两点基本假设,为了更好地解释本研究的内容,笔者对这两点假设进行一定修改后如下:①重要性高的实体比重要性低的实体更有用,更能表示文档的主要信息;②重要性越低的实体的排序越低,价值越低,因为这样的实体不是核心实体,未能代表文档的主要信息。

对于排在 n 位的实体,其 NDCG 的计算公式如下:

$$N(n) = Z_n \sum_{j=1}^{n} \frac{2^{r(j)} - 1}{\log(1+j)} \qquad 公式(11)$$

其中 Z_n 是指规范化因子,保证 N(n)2[0;1];r(j)是指第 j 个结果的重要性等级;$2^{r(j)}-1$ 是指第 j 个结果的贡献值,各类实体的重要性等级及其贡献值见表1;log(1+j)是指位置折扣,对数以 2 为底。对于错误识别的命名实体,标注者将实体的重要性等级标注为0,其贡献值也为0。

表 1　实体重要性等级及其贡献值

重要性	重要性等级	贡献值
核心实体	4	15 =2^4 －1
重要实体	3	7 =2^3 －1
弱相关实体	2	3 =2^2 －1
边缘实体	1	1 =2^1 －1

4.3.2　逆序对比率　假设 e_i 和 e_j 是新闻文档中的两个实体且实体 e_i 的重要性等级高于 e_j,理想情况下 e_i 排序在 e_j 之前,则 $<e_i, e_j>$ 是一个正序对。如果 e_i 排序在 e_j 之后,则 $<e_i, e_j>$ 是一个逆序对。逆序对比率是指排序结果中逆序对数目占标注数据中正序对数目的比例。此外,论文约定同一重要性等级的实体之间既不是正序对,也不是逆序对。

4.3.3　两个指标的关系　NDCG 和逆序对比率都可以用来评价实体重要性排序结果,二者在一定程度上呈正相关关系,但是这并不意味着二者存在线性关系。例如,对于同一篇文档,某种方法产生了一个核心实体与重要实体的逆序对,另一种方法产生了一个弱相关实体与边缘实体的逆序对,两个方法的逆序对数目相同,从而逆序对比率也相同。但是由于前一个逆序对的实体更加重要,其折扣的贡献值更大,故前者的 NDCG 值大于后者。

4.4　实验结果及分析

表 2 呈现了实体重要性排序实验的结果。从表 2 中可以看出,基于 TF＊IDF 的实体重要性判断方法,其 NDCG 值达到最大。其与基于实体频率的方法的区别在于,融合了实体的全局特征逆文档频率。相比之下,其效果有了提升,其中 NDCG 值提升了 2.71%,逆序对比率提升了 2.6%。而融合实体全局特征的集成方法也获得了较好的效果,该方法的逆序对比率值达到0.844 6,是所有方法中的最高值。

表 2　实体重要性排序实验结果

方法	NDCG	逆序对比率
实体频率	0.933 3	0.818 3
聚集系数	0.757 3	0.68
信息熵	0.926 9	0.801 5
TextRank 方法	0.915 8	0.806 2
TF＊IDF	0.958 6	0.839 6
集成方法	0.957 8	0.844 6

5　结语

本研究在已有研究的基础上通过利用基于实体频率、TF＊IDF、聚集系数、信息熵、TextRank 等方法以及集成方法,对文档的实体重要性进行排序实验。相比现有研究,本研究关注了一个相对较新的研究问题,即面向单文档的实体重要性排序,并对该问题进行了初步探索。本研究对搜狗全网新闻数据集进行处理,在此数据集上进行分类实验;定义了实体重要性等级并从数据集中随机抽取 50 篇进行实体重要性等级标注;利用 NDCG 和逆序对比率两个指标评价排序结果。实验结果表明,基于聚集系数的方法效果一般,而其他方法能获得较好的效果。基于聚集系数的方法假设聚集系数越大,实体越重要,然而,该假设并不总是成立。在某些文档中边缘实体的聚集系数很高,而核心实体的聚集系数较低。

尽管本研究的实验取得了一定的效果,但是也存在不足之处。本研究只考虑了人物、地点、机构等 3 类实体,但是某些类型的新闻文档可能不包含这 3 类实体,例如“健康”类的新闻。实体本身的概念比较广泛,既包括人名、地名、机构名等具体事物,还包括关系、概念等抽象事物。这些抽象实体的属性对处理新闻文档也具有重要意义。在今后的研究中需要引入更多的实体类型,使得研究更加合理、更具意义。此外,本研究的数据来自人工标注,可能存在主观偏差。理想情况下要尽可能地从用户生成内容中挖掘相关标记,一方面便于训练机器学习算法,另一方面也便于大规模测评,更好地提升效果。

参考文献:

[1] 张晓艳,王挺,陈火旺. 命名实体识别研究[J]. 计算机科学, 2005, 32(4):44－48.
[2] 陆伟,武川. 实体链接研究综述[J]. 情报学报, 2015(1):105－112.

[3] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2):1-6.
[4] LIU M, LIU Y, XIANG L, et al. Extracting key entities and significant events from online daily news[C]// LI T. Proceedings for the 9th international conference on intelligent data engineering and automated learning. Berlin: Springer-Verlag, 2008:201-209.
[5] TRANI S, LUCCHESE C, PEREGO R, et al. SEL: a unified algorithm for salient entity linking and saliency detection[C]// SABLATNIG R, HASSAN T. Proceedings for the 2016 ACM symposium on document engineering. New York: ACM, 2016:85-94.
[6] 郭艳卿,赵锐,孔祥维,等. 基于事件要素加权的新闻摘要提取方法[J]. 计算机科学, 2016(1):237-241.
[7] 吴玲达,雷震,老松杨,等. 基于局部话题句群的事件相关多文档摘要研究[J]. 计算机仿真, 2006, 23(11):263-267.
[8] KIRITOSHI K, MA Q. Named entity oriented related news ranking [M]. Berlin: Springer International Publishing, 2014:82-96.
[9] LIU M, LIU Y, XIANG L, et al. Single Chinese news article summarization based on ranking propagation[C]// 2008 International symposium on knowledge acquisition and modeling. Piscataway: IEEE, 2008:779-783.
[10] BEEFERMAN D, BERGER A, LAFFERTY J. A model of lexical attraction and repulsion [C]// COHEN P R, WAHLSTER W. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1997:373-380.
[11] NIESLER T. R., WOODLAND P. C. Modelling word-pair relations in a category-based language model[C]// 1997 IEEE international conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1997, 2: 795-798.
[12] 搜狐实验室. 全网新闻数据[EB/OL]. [2017-03-16]. http://download.labs.sogou.com/dl/.
[13] PANTEL P, FUXMAN A. Jigs and lures: associating web queries with structured entities [C]// LIN DK. Proceedings of the 49th annual meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2011:83-92.
[14] LIN T, ETZIONI O. Entity linking at web scale[C]//Proceedings of the joint workshop on automatic knowledge base construction and Web-scale knowledge extraction. Stroudsburg: Association for Computational Linguistics, 2012: 84-88.
[15] WELTY C, MURDOCK J W, KALYANPUR A, et al. A comparison of hard filters and soft evidence for answer typing in Watson[C]// Proceedings of the 11th international conference on the Semantic Web - volume part II. Berlin: Springer-Verlag, 2012:243-256.
[16] ZHANG H P, YU H K, XIONG D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]// Proceedings of the second SIGHAN workshop on Chinese language processing-volume 17. Stroudsburg: Association for Computational Linguistics, 2003:184-187.
[17] SHANNON C E, WEAVER W, WIENER N. The mathematical theory of communication[J]. Philosophical Review, 1949, 27(4): 623-656.
[18] CROFT W B. Search engines: information retrieval in practice [M]. METZLER D, STROHMAN T. 北京:机械工业出版社, 2009:1254-1271.

作者贡献说明:
陆娜:研究方案设计,具体实验,论文起草及修订;
周鹏程:研究方案设计和修订,论文修订;
武川:协助方案设计,参与论文修订。

Importance Based Entity Ranking for News Documents

Lu Na[1] Zhou Pengcheng[2] Wu Chuan[2]
[1] School of Information Science and Technology, Hainan Normal University, Haikou 571158
[2] School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] We propose an importance based method for entity ranking. Entities in a particular document show different importance. Many researches focus on documents or entities, such as text categorization and entity linking, while few research pay attention to the importance of entities in documents. This research has significant theoretical and practical value. [**Method/process**] Given a document which consists of words and entities, our method computes the relative importance of entities in the document, and then ranks these entities based on their importance with respect to the document. We perform experiment on the Sogou News dataset, and use evaluation metrics such as NDCG and inversed pair rate to evaluate the results. [**Result/conclusion**] Experimental results show that methods based on entity frequency, TF * IDF, distribution entropy and TextRank achieve better performance, while method based on cluster coefficient does not work well. In terms of NDCG, TF * IDF method reaches 95.86%, which is the best result and in terms of the inverse rate, the ensemble method reaches 84.46%, which is the best result.

Keywords: news documents entity importance entity ranking